How physicists analyze massive data: LHC + brain + ROOT = Higgs

Axel Naumann, CERN - axel@cern.ch

33C3, 2016 (but almost 2017)

CERN, People, Code

Axel Naumann, CERN - axel@cern.ch
33C3, 2016 (but almost 2017)

Content

- CERN
- How we do physics
- Computing
- Data
- Data analysis model in high energy physics
- Future of data analysis

CERN

"What is CERN" in 1 Minute

- European Organization for Nuclear (read: Particle!) Research, est. 1954, near Geneva
- Fundamental research (WWW: inventions happen)

```
knowledge CERN(money, curious_brains)
```

- What is mass? What's in the universe? Probing smallest scale particles: Higgs particle, super symmetry,...

Fact Sheet

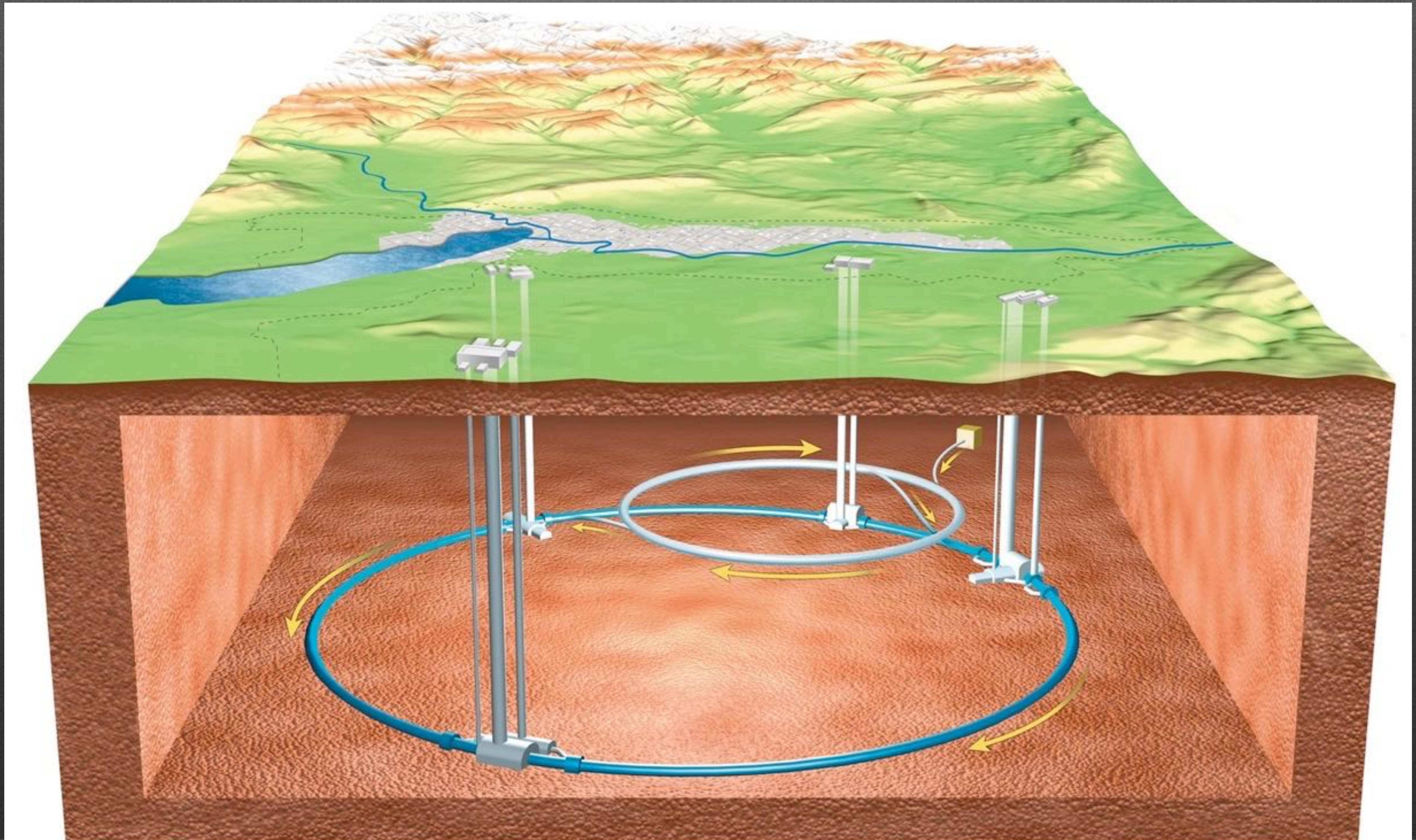
- CERN facilities used
  - by 12,000 physicists
  - from 120 nations
- CERN itself has approximately 2500 employees

Large Hadron Collider

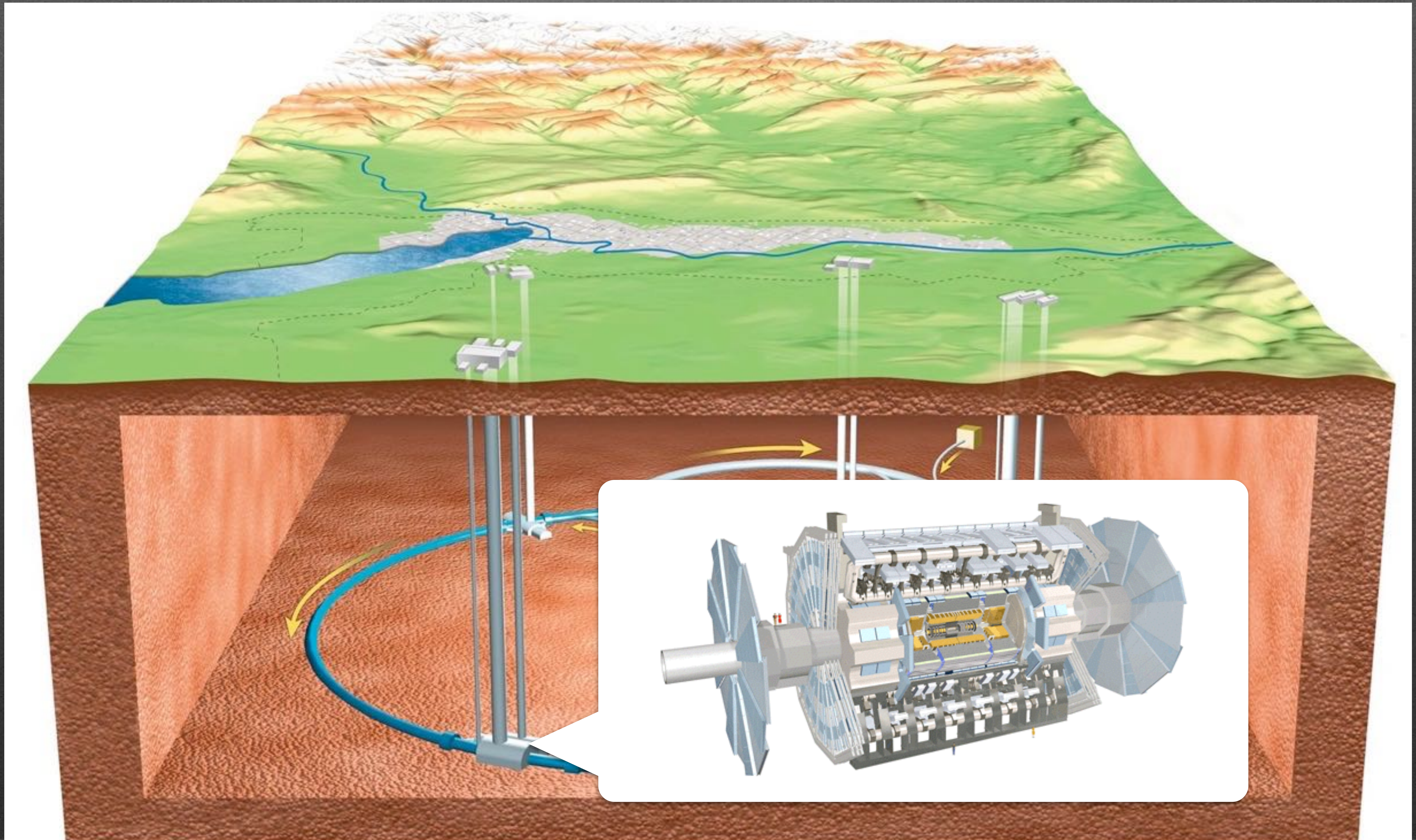
Large Hadron Collider


17250

Large Hadron Collider



Large Hadron Collider



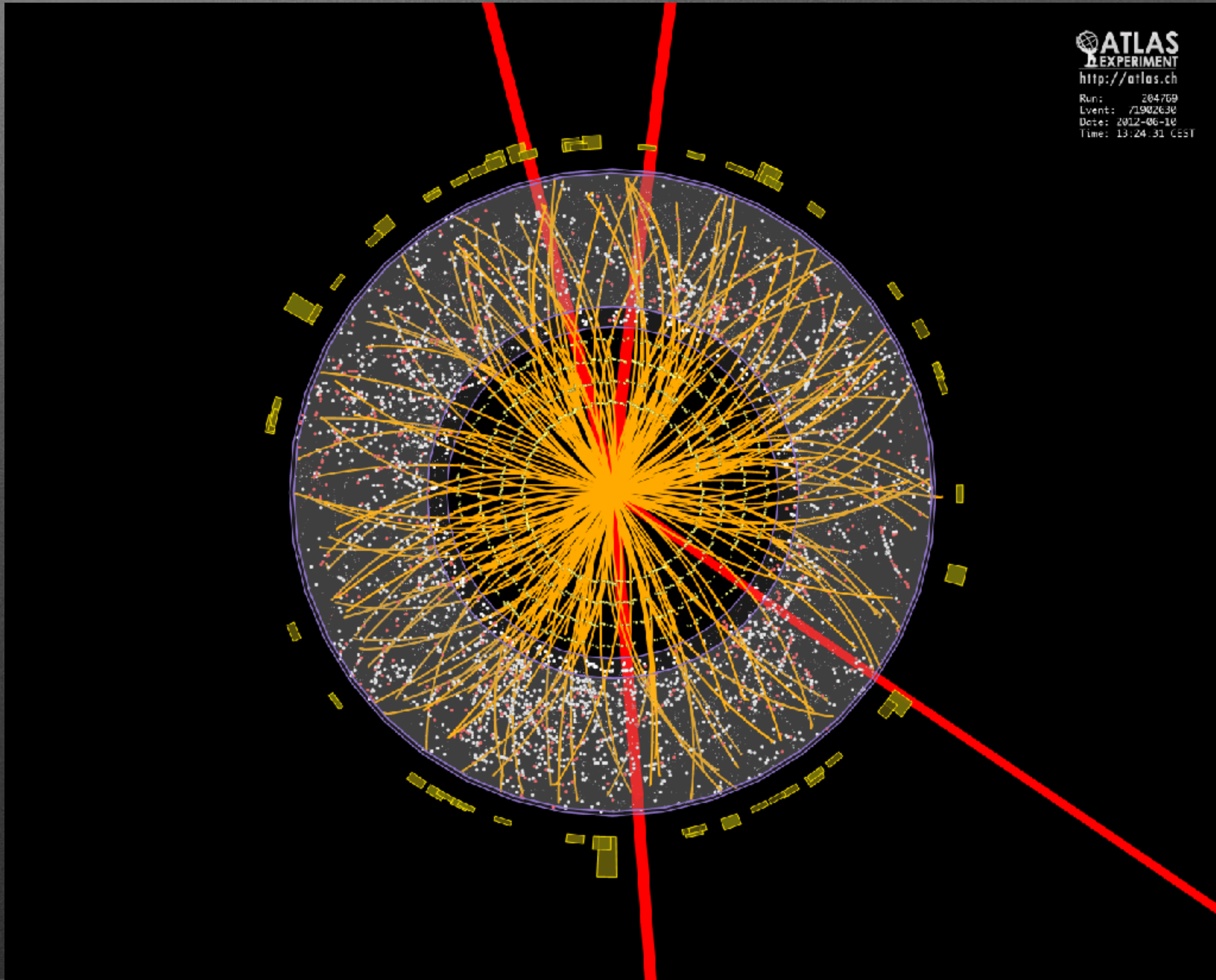
Large Hadron Collider

- World's "biggest" particle accelerator
- Ring with 27km in circumference, 100m below Switzerland and France
- Four large experiments ALICE, ATLAS, CMS, LHCb
- Expected to run until approximately 2030


ME&RE+1/3/04
ME&RE+1/2/02
ME&RE+1/2/01
ME&RE+1/2/36
ME&RE+1/2/34

Detectors

- Like a massive camera
    - O(100M) “pixels”
    - O(100M) pictures per second
- Identify particles
- Measure their properties



Life at CERN

Work At CERN

Data Taking

Studying the Forces

Scientific Discourse

Lecturing and Being Lectured

?!

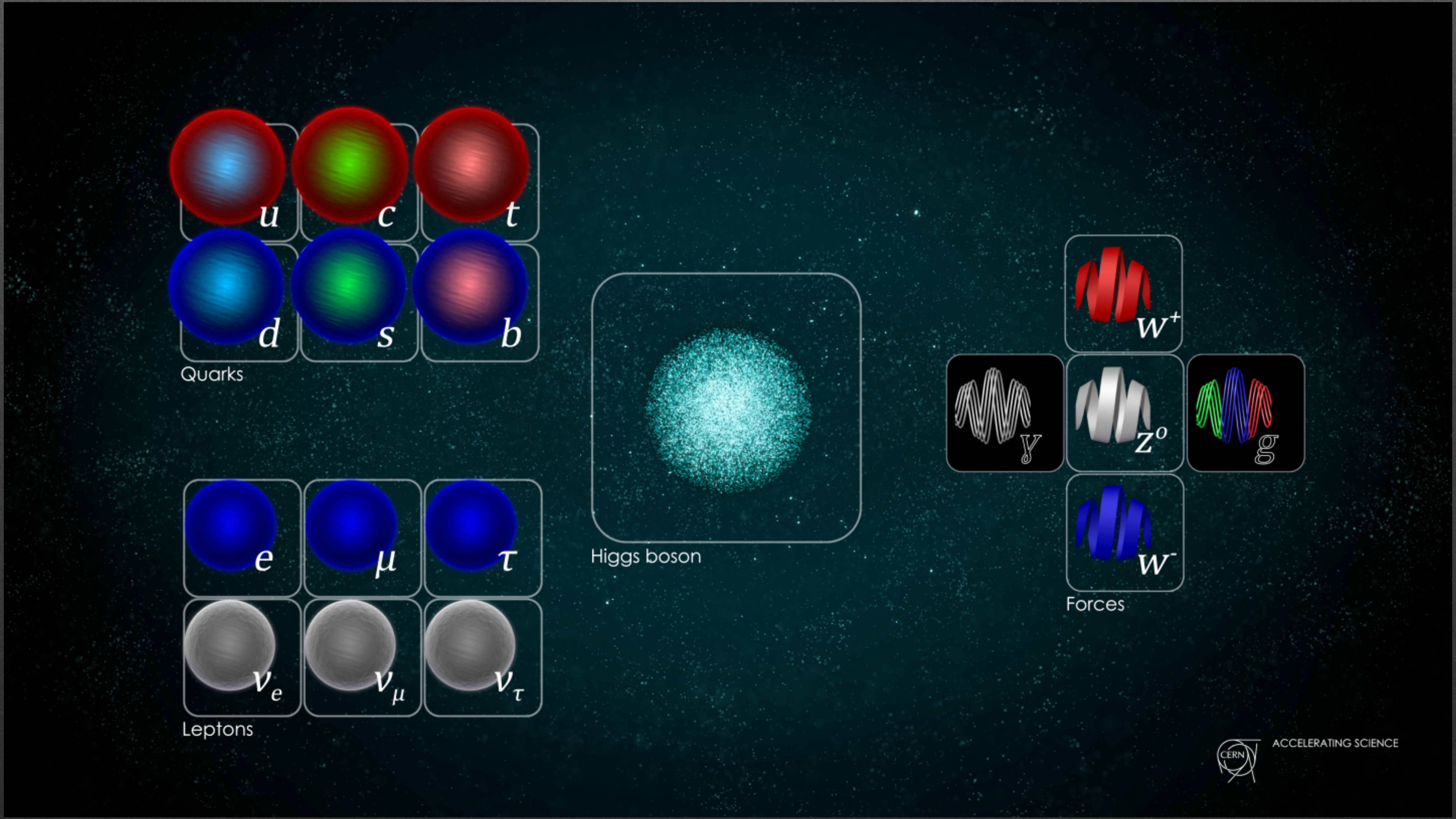
Presentational Democracy:
choose your own talk!

1) physics

2) model, simulation, data [p31]

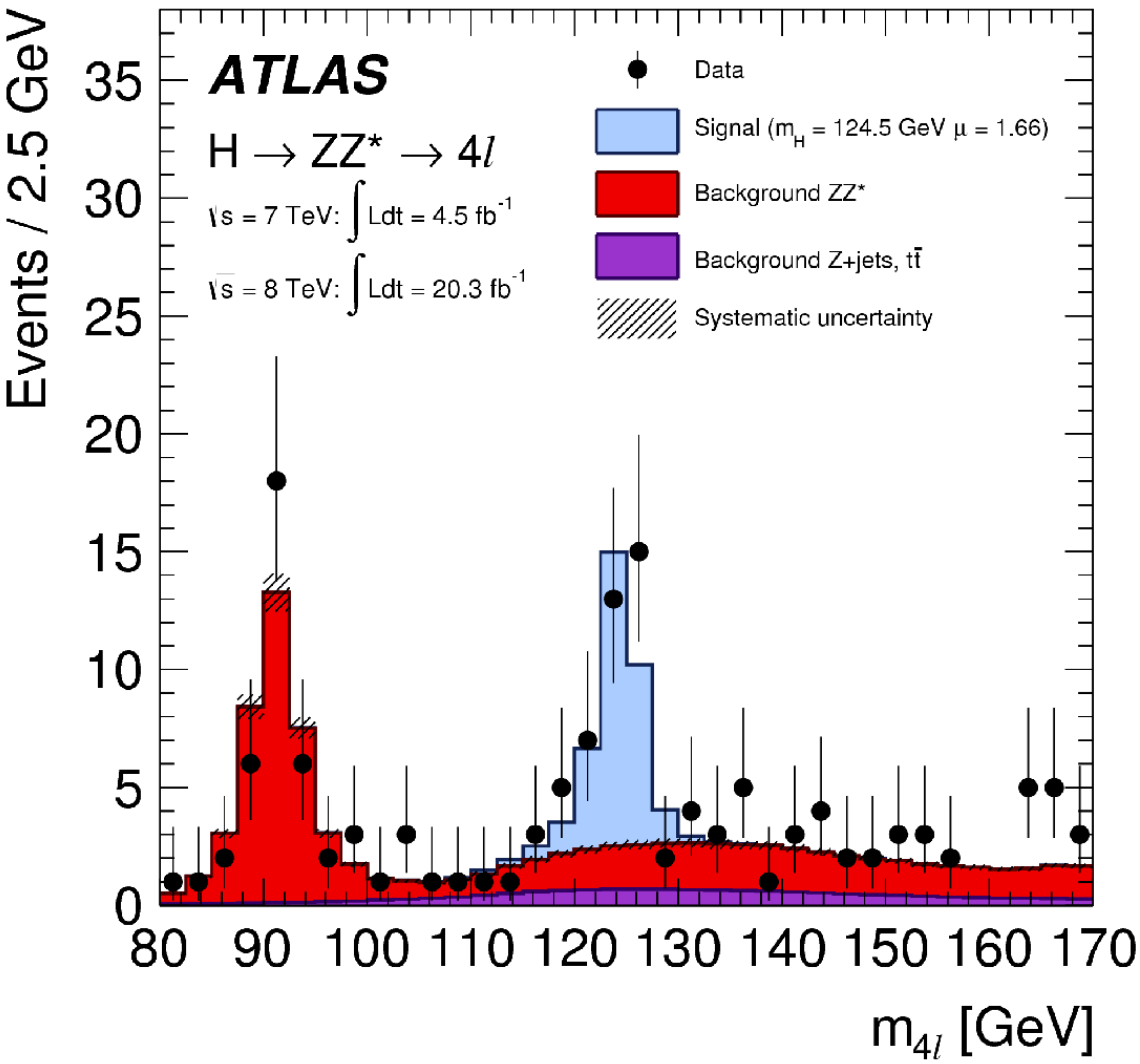
1)
We Do Physics. Here’s Yours.

Our Bits



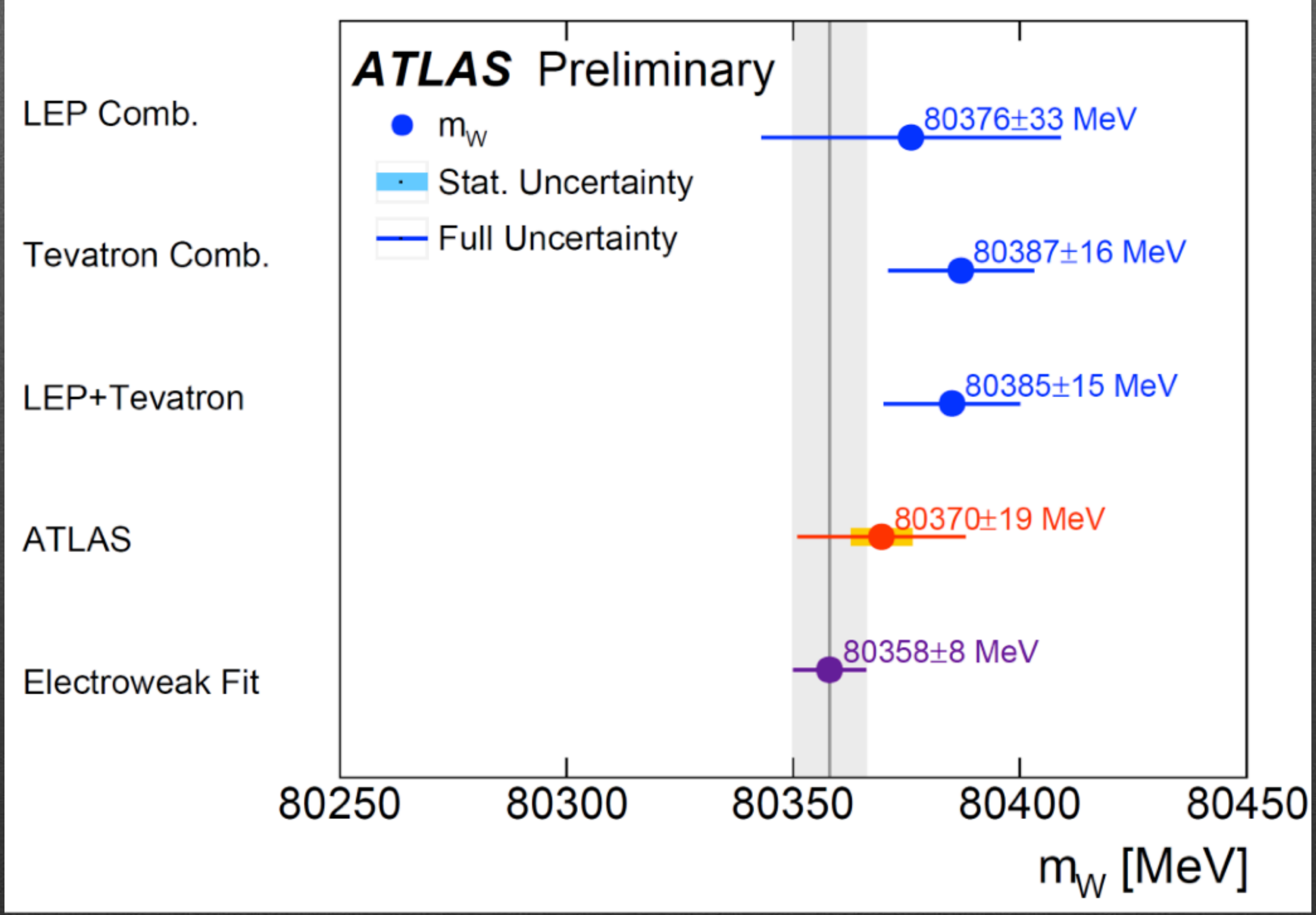
u
c
t
d
s
b
Quarks
e
μ
τ
ν_e
ν_μ
ν_τ
Leptons
Higgs boson
W^+
γ
Z^0
g
W^-
Forces
CERN
ACCELERATING SCIENCE

3 minutes
10−5 seconds
10−10 seconds
10−34 seconds
10−43 seconds
1032 degrees
1027 degrees
1015 degrees
1010 degrees
109 degrees
6000 degrees
radiation
particles
positron (anti-electron)
proton



ATLAS
H → ZZ* → 4l
√s = 7 TeV: ∫Ldt = 4.5 fb⁻¹
√s = 8 TeV: ∫Ldt = 20.3 fb⁻¹
Data
Signal (m_H = 124.5 GeV μ = 1.66)
Background ZZ*
Background Z+jets, tt̄
Systematic uncertainty
Events / 2.5 GeV
m_4l [GeV]
0
5
10
15
20
25
30
35
80
90
100
110
120
130
140
150
160
170

Newest Results



You survived.

[continue at p37]

2)
Model, Simulation, Data

Theory and Simulation

- Super *SUPER!!!1* precise
- But LHC experiments also looking for unconfirmed / weird things
  - monopoles, super symmetry, black holes, …
- Theory predicts production in collision, simulation predicts detector's view

Prediction versus Measurement

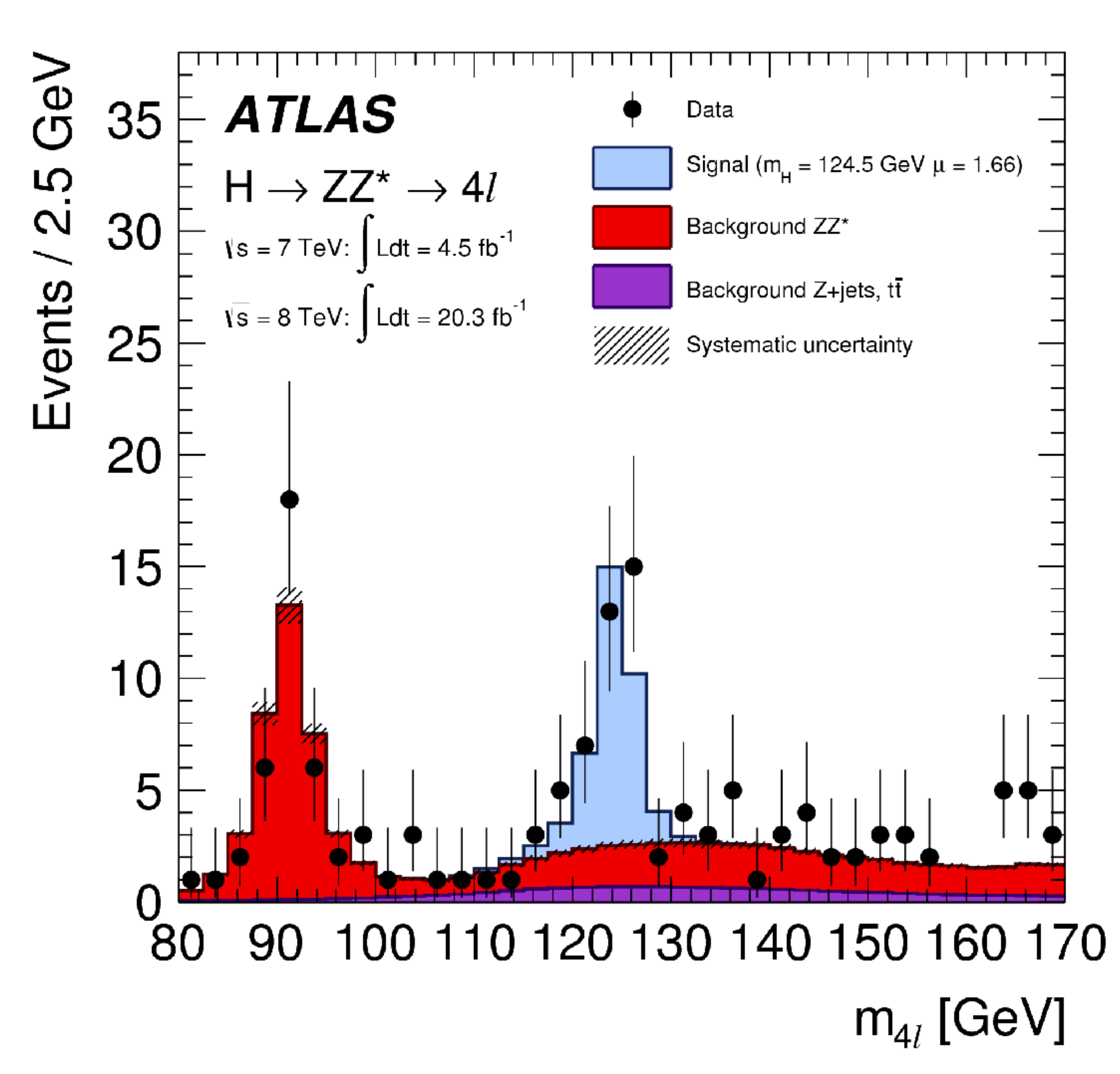
- When is a difference between “boring theory” and measurement significant enough to claim “this is new physics”?
  - detector simulation: how much do I expect?
  - reconstruction software: how much did I get?
  - statistics: is that expected?

Let’s Talk Weather versus Climate

- Measure temperatures
- Detect “abnormal” temperature variations (i.e. climate effects)
  - more measurement periods help
  - larger deviations help

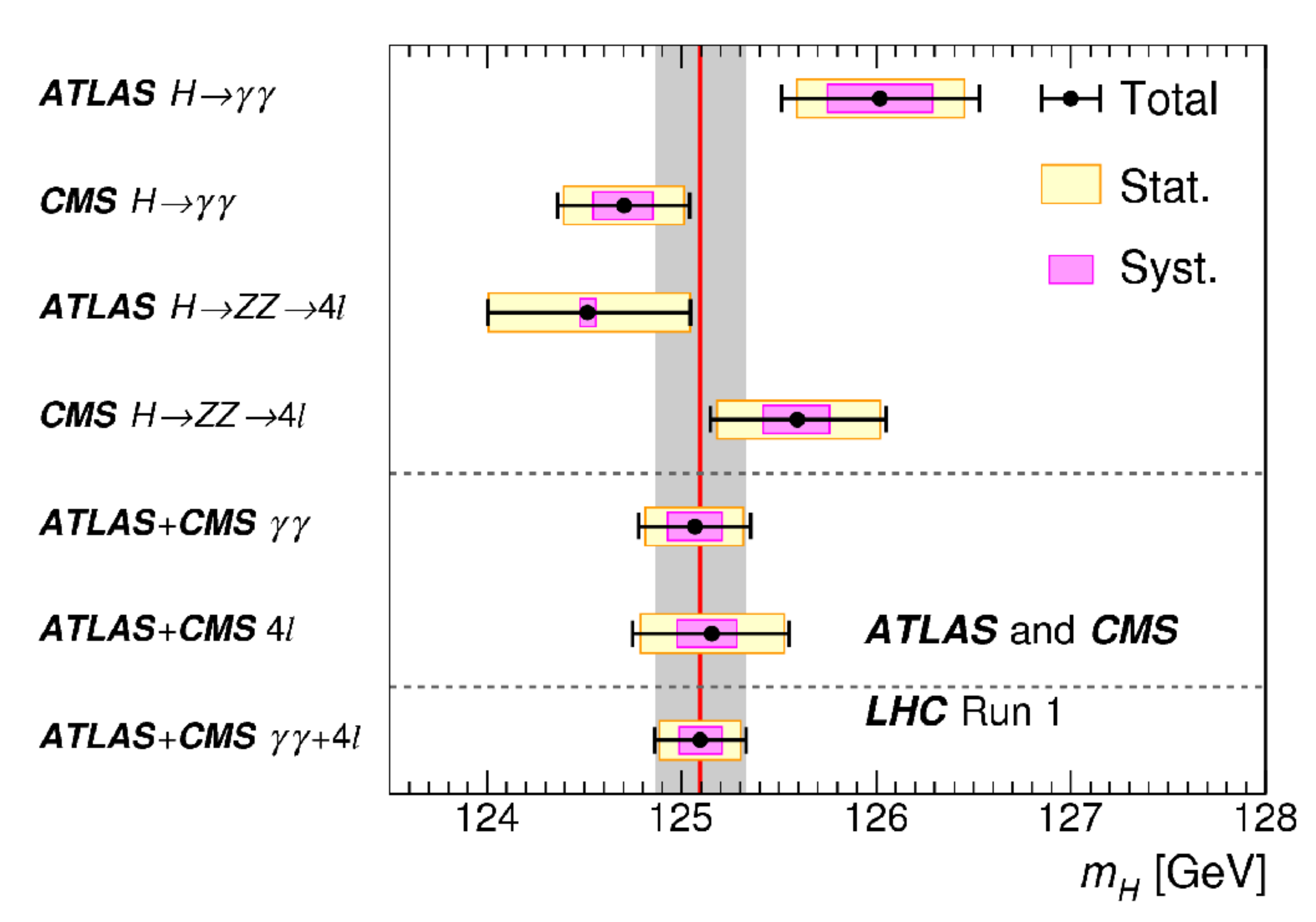
Data and Uncertainties

- Our simulation has uncertainties from theory
- Our measurements have uncertainties
- Our measurements have multiple contributions; need to track known versus new physics



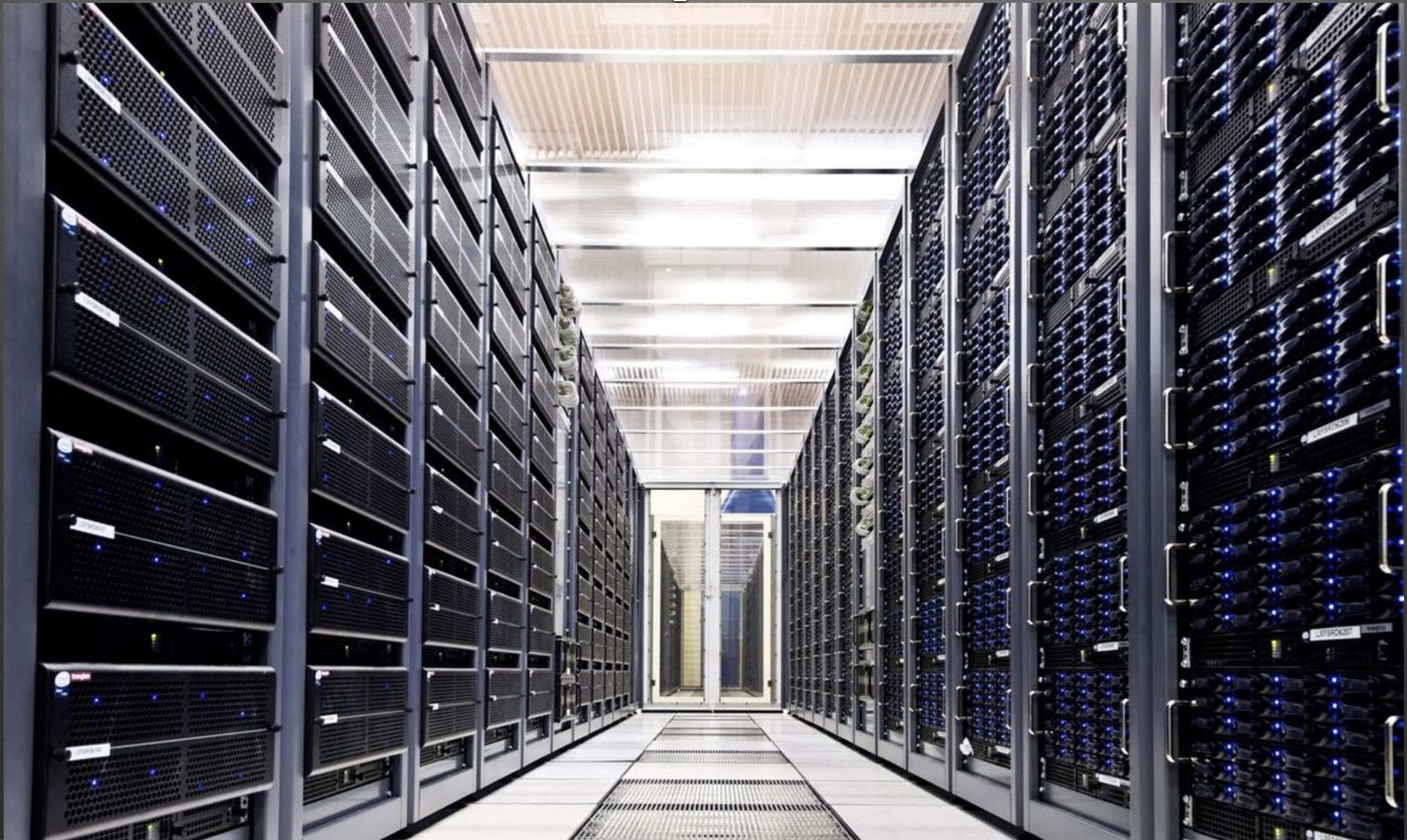
More Data Helps

- Correlating data helps
- Reduced measurement uncertainty helps
- More collisions = more data = higher chance to claim "we see something"

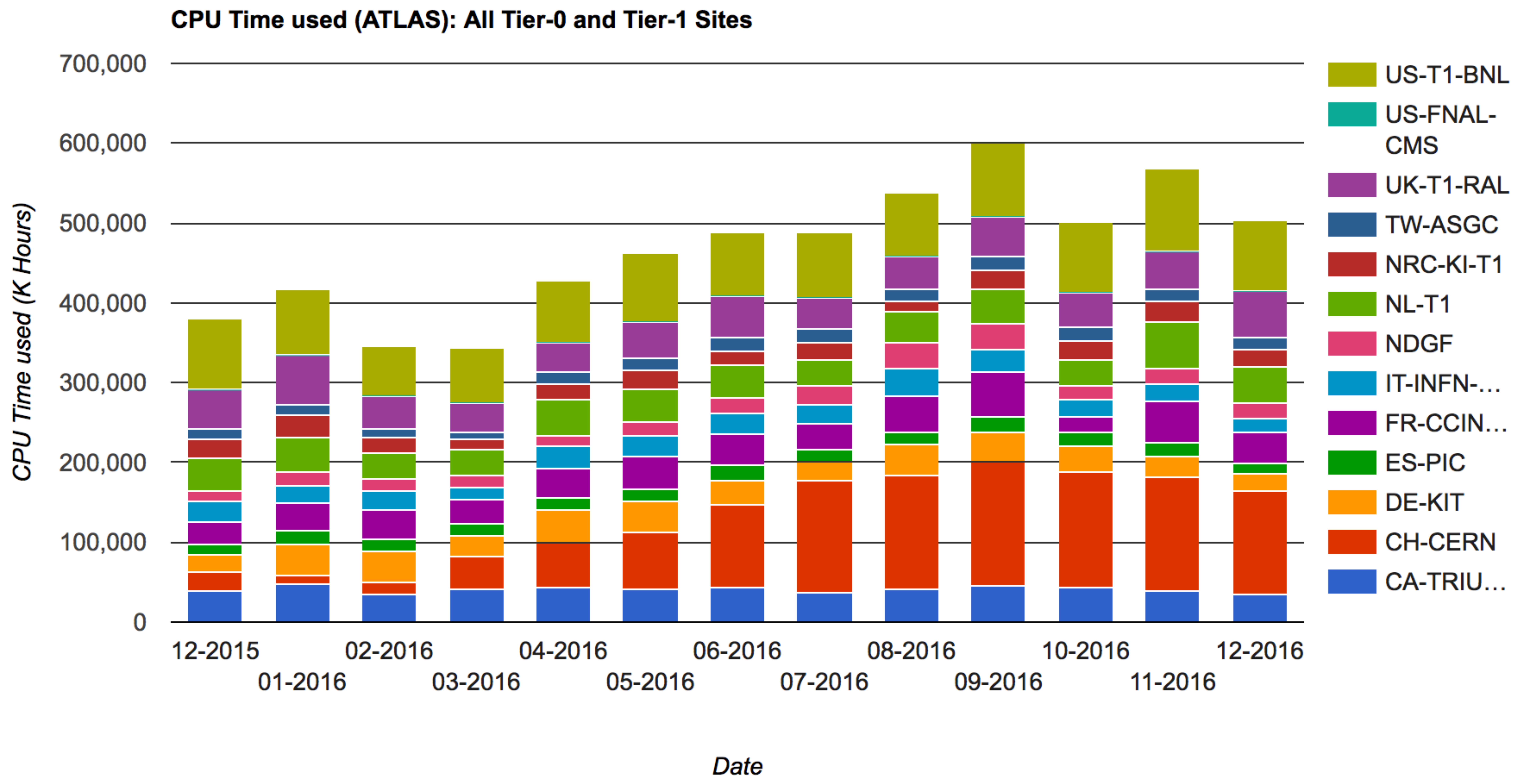


Computing

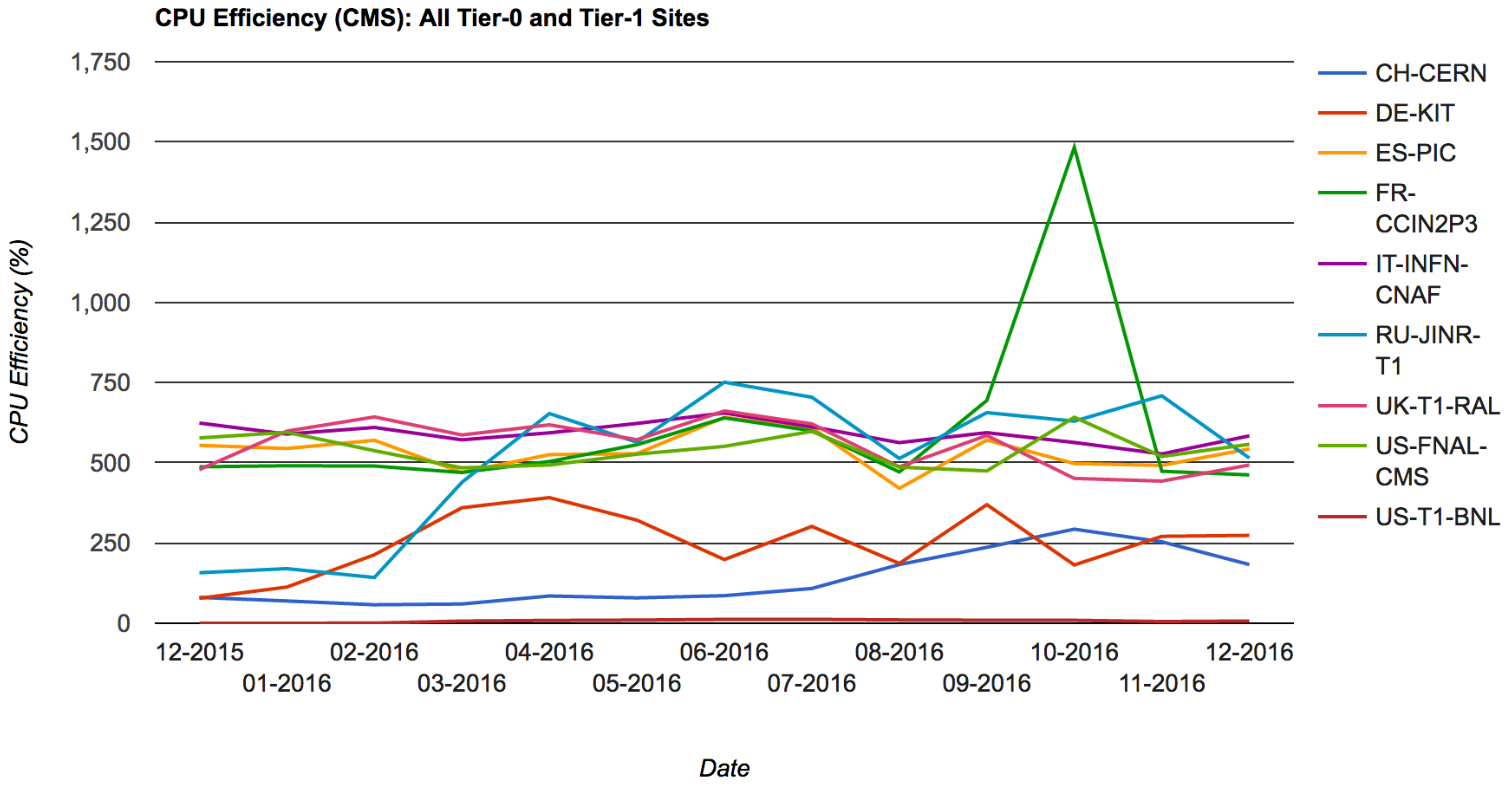
Computers



Computers



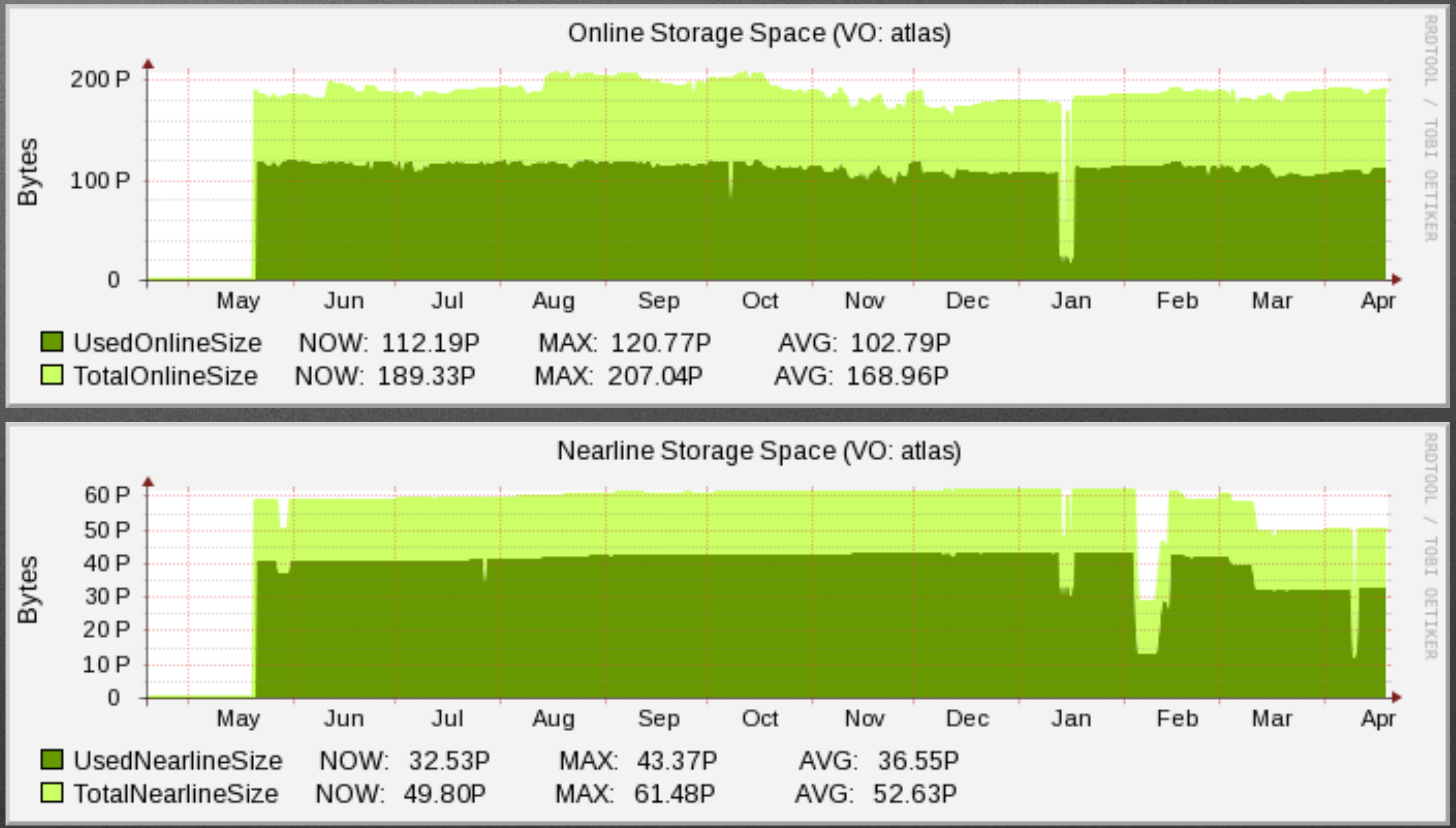
CPU Time used (ATLAS): All Tier-0 and Tier-1 Sites
CPU Time used (K Hours)
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
12-2015
01-2016
02-2016
03-2016
04-2016
05-2016
06-2016
07-2016
08-2016
09-2016
10-2016
11-2016
12-2016
Date
US-T1-BNL
US-FNAL-CMS
UK-T1-RAL
TW-ASGC
NRC-KI-T1
NL-T1
NDGF
IT-INFN-...
FR-CCIN...
ES-PIC
DE-KIT
CH-CERN
CA-TRIU...



CPU Efficiency (CMS): All Tier-0 and Tier-1 Sites
CPU Efficiency (%)
1,750
1,500
1,250
1,000
750
500
250
0
12-2015
01-2016
02-2016
03-2016
04-2016
05-2016
06-2016
07-2016
08-2016
09-2016
10-2016
11-2016
12-2016
Date
CH-CERN
DE-KIT
ES-PIC
FR-CCIN2P3
IT-INFN-CNAF
RU-JINR-T1
UK-T1-RAL
US-FNAL-CMS
US-T1-BNL

Storage

- Tera, Peta, Exa:
  1EB = 1,000,000 TB
- Capacity: 0.7 EB
- Usage: 0.7 EB



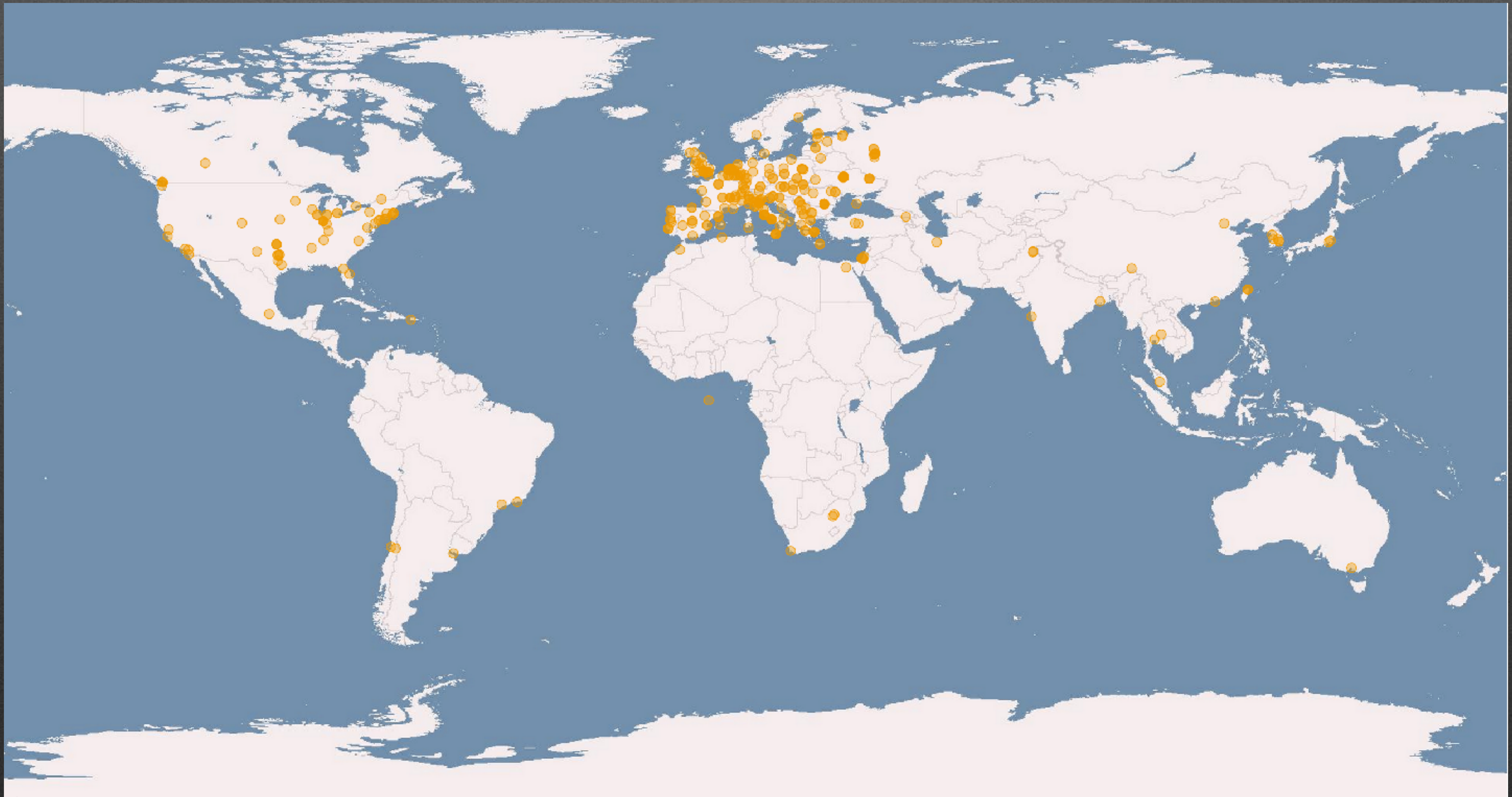
End 2015, before new data taking run

?!

1) distributed computing
2) measure effects of bugs [p50]

1)
Distributed Computing

170 Compute Centers
= The Grid



The Grid

- “WLCG”: world-wide LHC computing grid
- About 600,000 cores
- Used for large-scale data operations

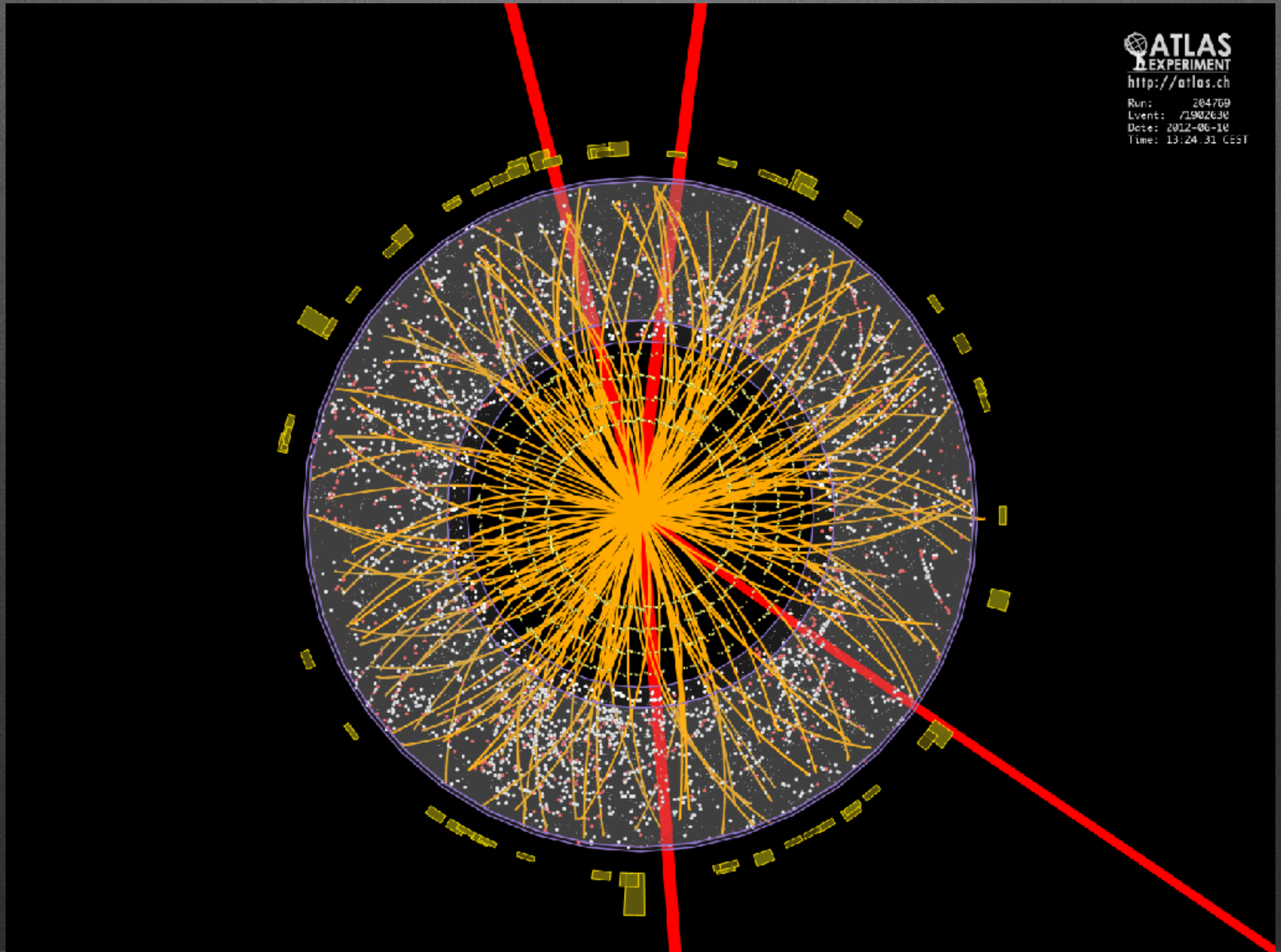
Why?!

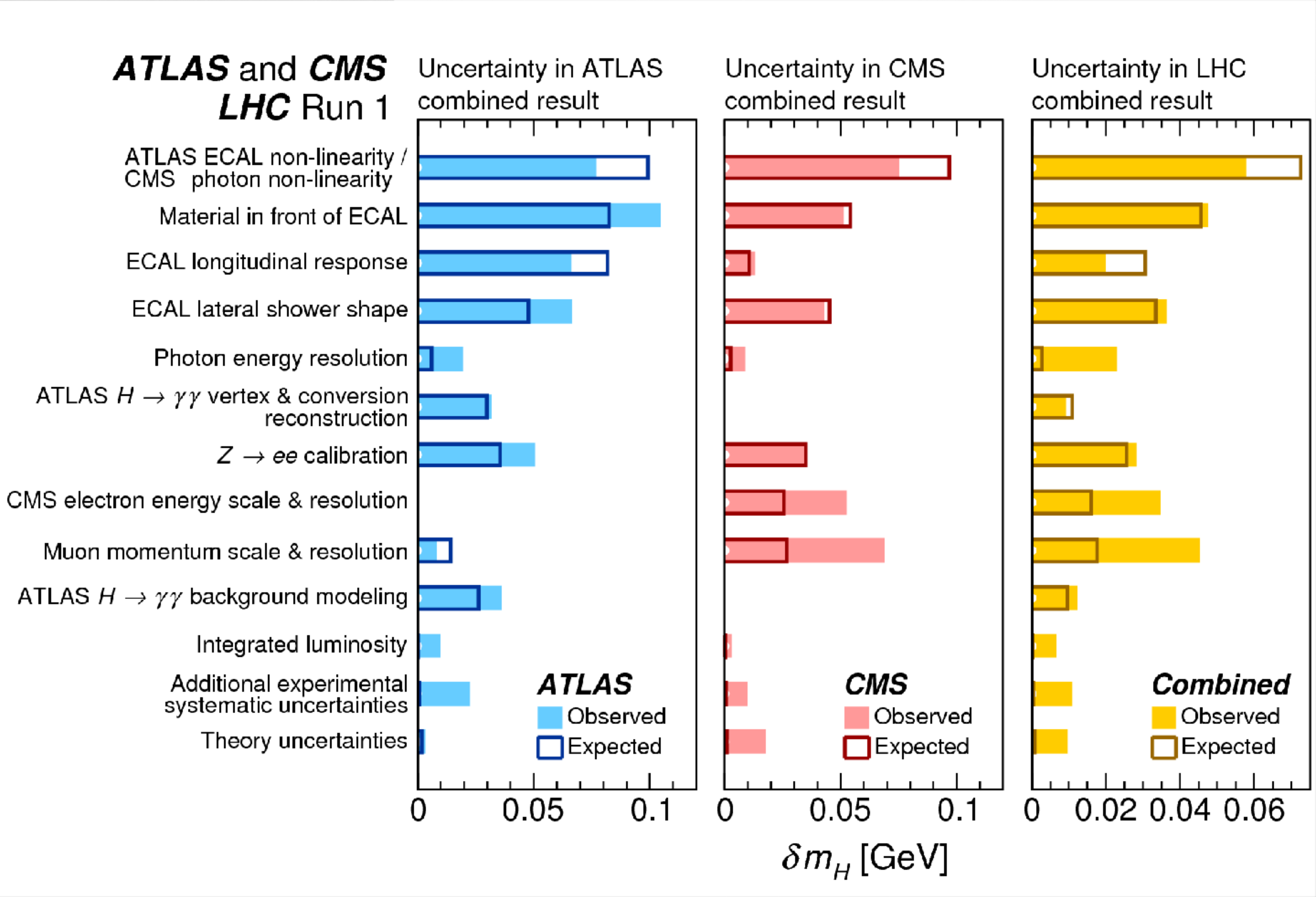
- Easier to get countries to commit domestic resources
- (Claim?) synergy with other sciences. Today we’d call it “cloud”
- But underestimated cost (nerves + €/₽/¥/$/CHF) and data distribution issue
  - there’s always a holiday somewhere. And some universities’ summer vacation is brutal for operations.
- Still it just works, allows us to scale

[continue at p53]

2)
Measure Effects of Bugs

Bug?


ATLAS
EXPERIMENT
http://atlas.ch
Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST



ATLAS and CMS
LHC Run 1
Uncertainty in ATLAS combined result
Uncertainty in CMS combined result
Uncertainty in LHC combined result
ATLAS ECAL non-linearity / CMS photon non-linearity
Material in front of ECAL
ECAL longitudinal response
ECAL lateral shower shape
Photon energy resolution
ATLAS $H \to \gamma\gamma$ vertex & conversion reconstruction
$Z \to ee$ calibration
CMS electron energy scale & resolution
Muon momentum scale & resolution
ATLAS $H \to \gamma\gamma$ background modeling
Integrated luminosity
Additional experimental systematic uncertainties
Theory uncertainties
ATLAS
Observed
Expected
CMS
Observed
Expected
Combined
Observed
Expected
0
0.05
0.1
0
0.05
0.1
0
0.02
0.04
0.06
δm_H [GeV]

Data

./findHiggs --help

- Reconstruction done by multi-GB C++ programs
  - approx 50 millions lines of C++ at CERN
- Experiment-specific
  - centrally curated by experiments, e.g. http://cms-sw.github.io/
  - correct! efficient! Experiment decides what to spend CPU cycles on

2016Data.csv?

- Data in custom, binary format, since 20 years:
  ROOT files https://root.cern
  - collisions are (mostly) independent: can use “rows”
  - but nested collections, custom float precision
- Generated from C++ object layout (aka class definitions)
- Can be read in C++ as well as JavaScript, Scala (without C++ involved)

Why not MyPostacle?

- Databases etc didn’t (and don’t) scale
- Have C++ on reading and writing side
  - databases are a medium change
- Need only single collision’s data in memory!
- Concept “file system” is well understood, modeled, supported; it scales, is future-proof etc.

Why Not HDFS / HDF5 / Protobuf / …

- Want builtin schema evolution: changes in class layout of written data versus binary (requires layout to be stored as part of data)
- Want I/O without having to annotate / change code; automatic I/O from class definition instead of everyone writing serializers (and bugs)
- Rationale: robustness
  - besides brains, this data is our fortune. Must. Not. Lose.

?!

1) cling, our C++ Interpreter
2) Open Data and Applied Science [p70]

1)
cling, Our C++ Interpreter

WTF?!

```
****************** CLING ******************
* Type C++ code and press enter to run it *
*             Type .q to exit             *
*******************************************
[cling]$ #include <cmath>
[cling]$ std::sin(0.42)
(double) 0.407760
[cling]$ CMS::GetTheAnswer()
(int) 42
[cling]$
```

Exploring Code Through Experiments!

- We LOVE experiments.
- Did you ever probe functions using gdb?
- We use a C++ Interpreter: load complex parts, pick interfaces you need (or might need), test drive them!
- No linking, re-linking, and linking again
- Just keep trying (and keep saving - it *is* C++)

Explorative Coding

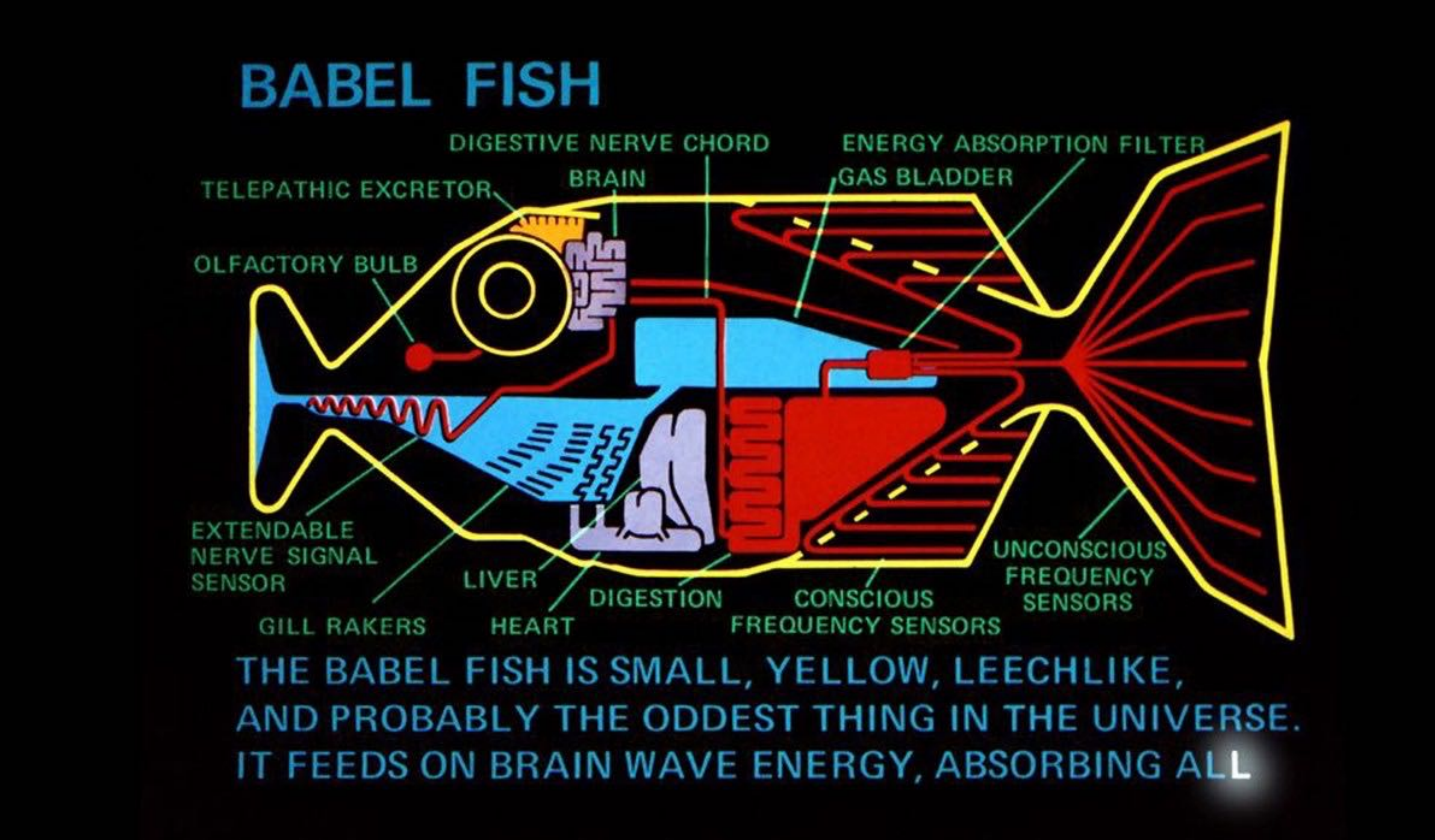
- Completely changed the way we (and especially novices) develop C++ code
  - organized framework used by creative, spontaneous, vivid scratch pad of all kinds of code
  - can shift from the scratch pad into the framework as code becomes stable and useful

Interpreting C++

- CINT from 1993-2013, based on the amazing Masaharu Goto
- Now cling cern.ch/cling based on clang + llvm
    - complete C++ support! Load libraries into cling, #include headers and hack away
    - see the unbiased e.g. https://youtu.be/BrjV1ZgYbbA

Under the Hood

- clang as C++ front-end
- llvm just-in-time compiles into memory
- need extensions / hacks to make “sin(0.42)” useful
  - expressions need to be executed
  - concept of “end of translation unit” is different for an interpreter
- Nonetheless, result is really natural



The Hitchhikers' Guide To The Galaxy

And Python, Too

- Do. Not. Use. SWIG. At least not on this scale.
- cling and Python share knowledge:
  - dynamic binding to Python, back and forth
  - C++ types in Python, C++ objects in Python!
  - Pythonization of C++ types: begin() + end()? iterable!
- Dynamic! At runtime! (Remember the vivid bubble?)

Interpreter + A Few = Serialization

- Interpreter governs AST
  - authoritative source of runtime reflection
- Build serialization on top
- Nicely scales to 0.x Exabytes of data, so far.

[continue at p86]

2)
Open Data, Applied Science

Budget

- 1.1B CHF = 1.0B EUR = 0.9B GPB = 1.1B USD
  - contribution by status, gross national product
  - Wikipedia: 2.2CHF / citizen / year
- THANK YOU.
- And: CONGRATULATIONS!

Society and CERN

- We can do what we do because of YOU
- We try to make EVERYTHING accessible to YOU
  - research results, in lots of forms
  - hardware
  - data
  - software

Sharing Research

- All publications Open Access, e.g. scoap3.org
    - a revolution!
- Immense effort goes into communication and “popularization”
    - we love to talk about what we do, we owe it to you to share, explain and answer what we can
- https://visit.cern/ - come visit us! (Pro tip: ask for underground tours by April!)

Applied Research

- Influence of cosmic rays in cloud formation
    - http://cern.ch/cloud
- Energy from nuclear waste
    - http://cern.ch/go/N7PL
- Re-purposing detectors
    - e.g. http://cern.ch/MEDIPIX

Hardware, Data,…

- Open Hardware www.ohwr.org/
  - e.g. White Rabbit: deterministic Ethernet
- Open Data opendata.cern.ch/
- LHC@home lhcathome.web.cern.ch/
  - and the new & excellent Virtual Atom Smasher test4theory.cern.ch/vas/

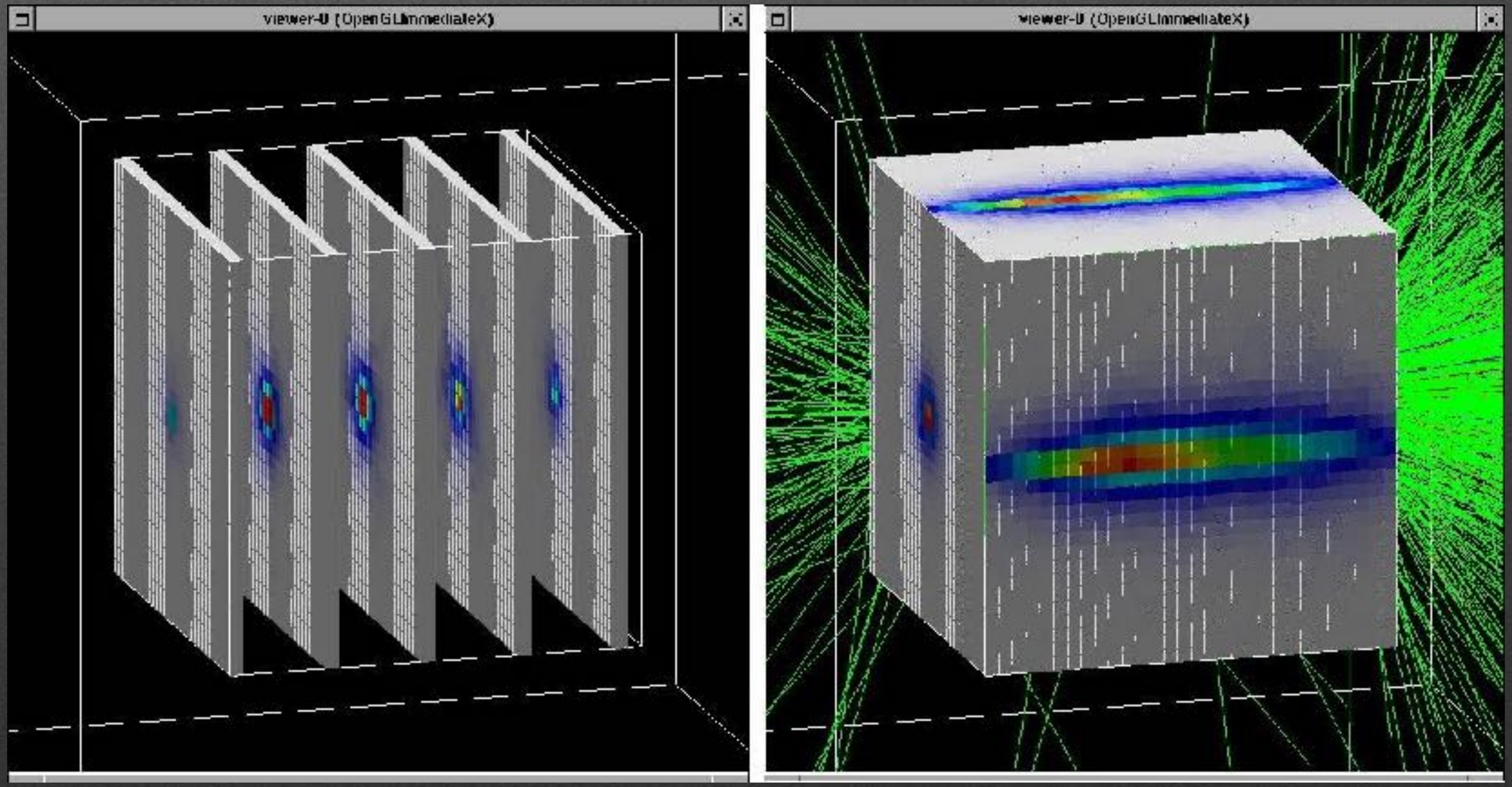
Using Open Source

- Almost everything at CERN is Open Source
- Use and contribute
    - GCC, clang, Puppet, OpenStack, Xen, Ceph, Jenkins, Andrew File System, LaTeX, Drupal,…

Creating Open Source

Geant

- Simulates interaction of particles with matter
    - used by people like us
    - NASA
    - medical radiation facilities
- geant4.cern.ch/



Indico

- Used to organize meetings and conferences
    - meeting room registration / search
    - manages time table, material, even paper reviewing
- Scales, production grade
    - > 20,000 users; protection / access schemes etc
- indico.github.io

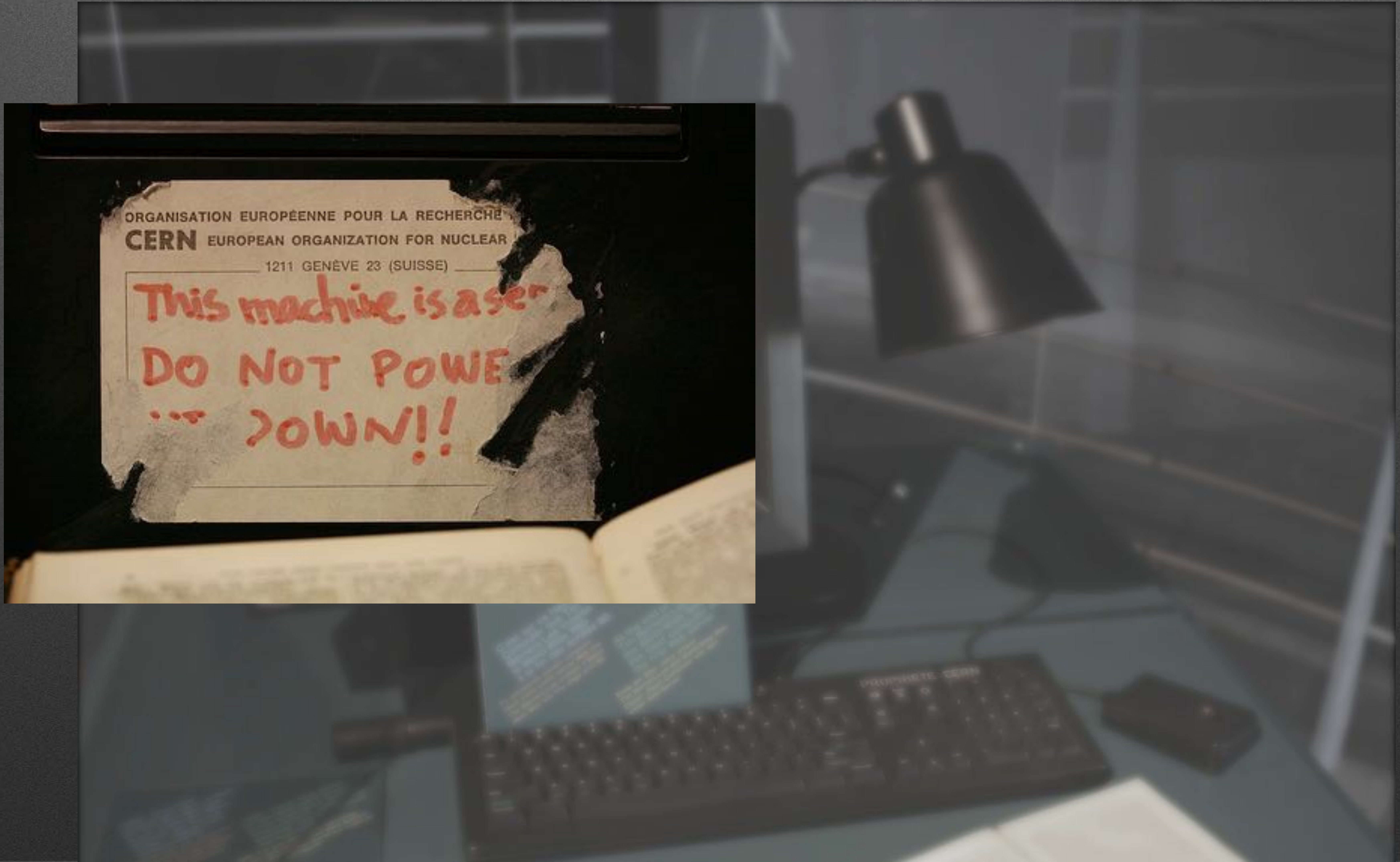
DaviX

- We love http!

WWW @ CERN


C'est sur ce Next Computer que Tim Berners-Lee inventa le World-Wide Web à la fin des années 80.
At the end of the 80s, Tim Berners-Lee invented the World-Wide Web using this Next computer as the first Web server.
PROPRIETE CERN

WWW @ CERN


ORGANISATION EUROPÉENNE POUR LA RECHERCHE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR
1211 GENÈVE 23 (SUISSE)
This machine is a se
DO NOT POWE
DOWN!!

DaviX

- We love http!
- Library for transparent http, WebDAV, S3 data transfer
- High throughput!
- Handles large collections of files
- cern.ch/davix

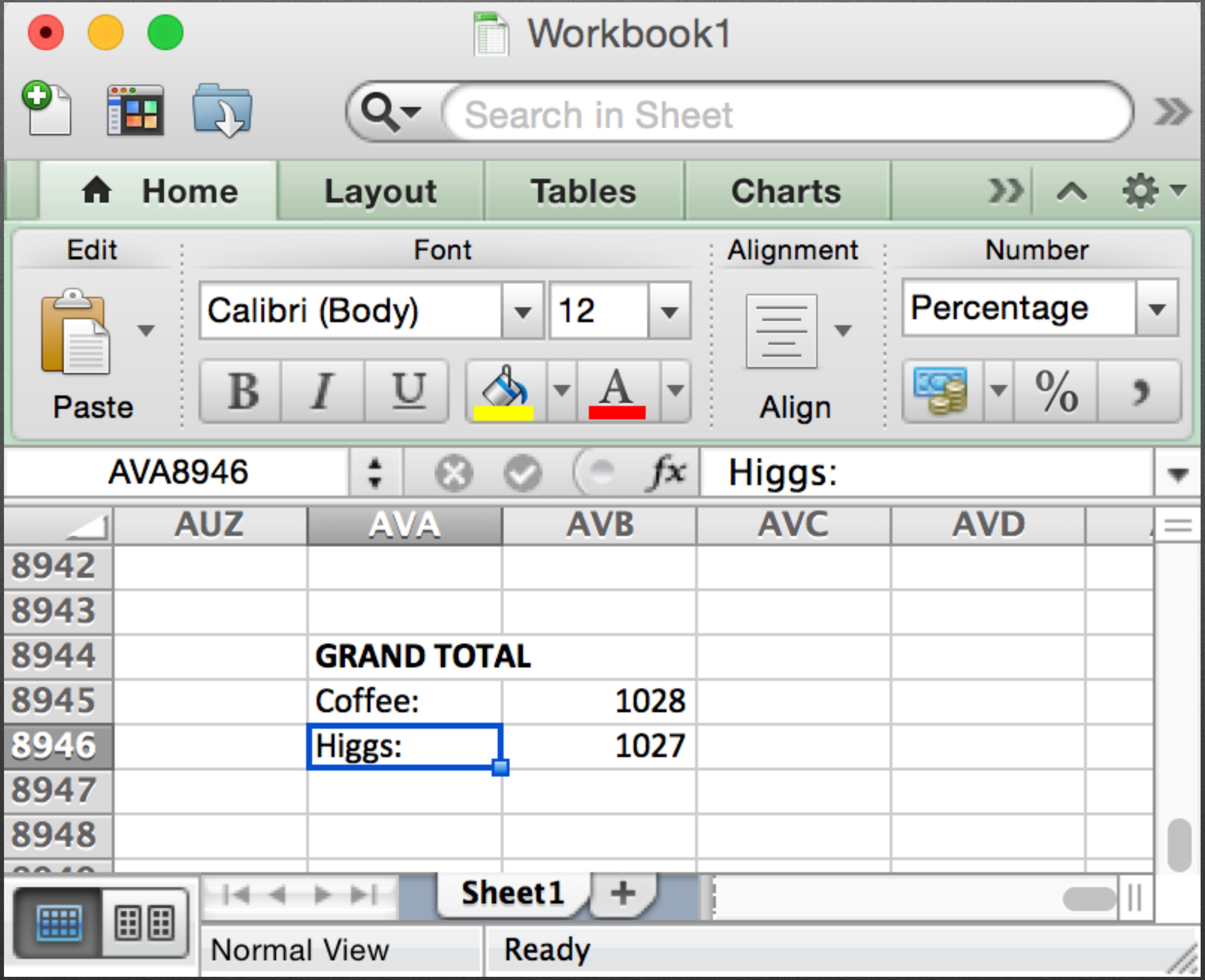
CernVM-FS

- Distribute huge releases onto 100,000 boxes: scp?
- No: cern.ch/cvmfs
  - http-based (!) network file system; write-few-read-many; robust, scalable
  - aggressive caching (even content-delivery systems)
  - can even boot a Virtual Machine out of thin air (but not vacuum)

ROOT

- Coming up…

Data Analysis in High Energy Physics


Workbook1
Search in Sheet
Home
Layout
Tables
Charts
Edit
Font
Alignment
Number
Calibri (Body)
12
Percentage
Paste
Align
AVA8946
Higgs:
AUZ
AVA
AVB
AVC
AVD
8942
8943
8944
8945
8946
8947
8948
GRAND TOTAL
Coffee:
1028
Higgs:
1027
Sheet1
Normal View
Ready

C++!

- Approx 50 million lines of C++ at CERN
- Very few devs have formal education in computer science / engineering
- C++ instead of Excel
  - Physicists write their analysis in C++! Themselves!

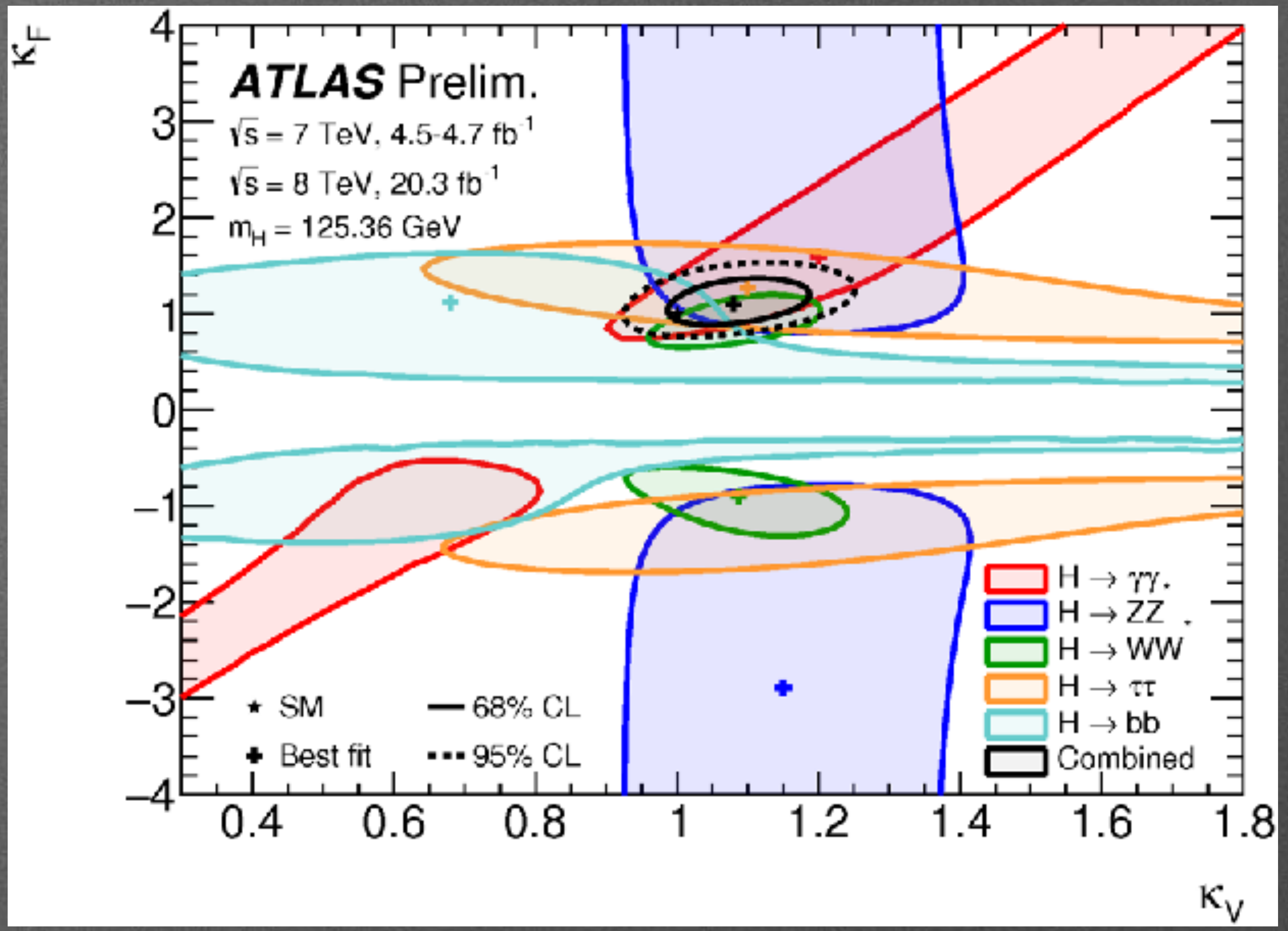
Key Features

- Keep only one collision in memory
- Throughput counts:
    - collisions / second
- Can specialize data format to optimize for specific physics analyses

ROOT

- root.cern
- Data analysis workhorse for all High Energy Physics
- Since 1995, now 2.5MLOC C++
- Physicists’ interface to huge, complex frameworks

ROOT



- Serialization facilities
- Statistics tools: modeling, determination of significance, multivariate
- Graphics to communicate results
- …which is all open source.
  - (… and guess who else is using it.)

By Allan Ajifo - CC BY 2.0

Conclusion

CERN and Society

- You enable great stuff - thank you!
- We want to share, and we do
  - we have good outreach people for science
  - not so much for software
  - but we do have good software! :-)

Scientific Computing

- Many building blocks existed outside our field, some crucial ones did not
  - C++ data serialization and distribution
  - efficient computing for non-computer scientists
  - scale, scale, scale
- More natural sciences arrive at the petabyte data range; they meet similar challenges

Forecasting Analyses: Characteristics

- Computing matters more an more: correlations become more important than data size
  - I/O (going random access for correlations) and CPU limitation
  - MVA still rising, but ends as a stat tool (except for generative part!)

Forecasting Analyses: Consequences

- Backend language matters: close to the metal, defines performance
- Design of analyses will generally not be graphics-based (“visual coding”) due to complexity
- Instead, need simple programming layer / different language: bindings matter!
- I/O must be adapted to analysis flow for max performance, e.g. "all data in memory" doesn't scale
- Throughput is king (think ReactiveX)

Contact

- Still here until tomorrow
- axel@cern.ch / Twitter's @n_axel_n